CROSS-SPECIES MICROARRAYS

The present invention relates to a method of identifying one or more oligonucleotides on a high-density oligonucleotide array (microarray) derived from a first species which can be used to analyse a corresponding nucleotide sequence in a second species, or a different variety of the first species; and the use of the identified oligonucleotides. Preferably if the array is used with a second species this species is related to the first species.

High-density oligonucleotide microarrays (WO9710365, EP0853679, US6040138, DE6962590T) are a powerful and commonly used tool for large-scale gene-expression profiling in organisms for which complete or extensive genome sequence data is available. A widely-used oligonucleotide microarray is based on GeneChip® technology (Affymetrix, Santa Clara, USA). GeneChip® microarrays use probe sets, comprising between 11 and 20 probe pairs, to quantify the abundance of each transcript. Each probe pair consists of a perfect match (PM) and a mismatch (MM) probe. The PM probe, which represents an identical copy of the target sequence, is typically a 25-base sequence complementary to the target transcript, whilst the MM probe is identical to the PM probe except for a single mismatch normally at the 13th base. The MM probes are designed to measure non-specific binding. Transcript abundance is calculated from hybridisation differences between PM and MM probes across a probe set. GeneChip® microarrays provide reproducible, accurate data at high throughput rates, and are the preferred option to provide the type of data for use within microarray databases where data standardisation is crucial. The wide-scale adoption of GeneChip® technology is exemplified within the plant sciences research community, where data from a large numbers of studies using *Arabidopsis thaliana* (L.) Heynh is publicly-available. These studies include plant responses to

biotic and abiotic stresses and comparisons between mutant or transgenic plants and their wild-types. At present GeneChip® microarrays are available for only a few species of eukaryotes. Thus, in contrast to *A. thaliana*, the study of the transcriptome for most agriculturally, or ecologically, important plant species has been inhibited since extensive sequence information and the fabrication of custom microarrays has been required.

According to an aspect of the invention there is provided a method for the identification of one or more oligonucleotides on a microarray derived from a first species which can be used to analyse a corresponding nucleotide sequence from a second species or a distinct variety of the first species, wherein the method comprises applying genomic DNA from the second species, or distinct variety of the first species, to the microarray derived from the first species, and identifying oligonucleotides on the microarray which hybridise to the genomic DNA.

The oligonucleotides on the microarray may also be referred to as probes, and the two terms are used interchangeably herein.

The first species may also be referred to as the reference species, and accordingly, the microarray derived from the first species may also be referred to as the reference species microarray.

A variety in the context of this invention may represent an ecotype or field accession, a strain or distinct laboratory subspecies, a commercial line, a 'sport' or mutant line, a natural or synthetic hybrid, a transgenic line, a breed or any combination of these types.

The corresponding nucleotide sequence in the second species, or the distinct variety of the first species, may be a part of a larger sequence.

The corresponding nucleic acid sequence is preferably a nucleic acid or nucleic acid sequence that is complementary, or substantially complementary, to the oligonucleotide and which can therefore under the appropriate conditions hybridise to the oligonucleotide. The nucleic acid may be deoxyribonucleotide or a ribonucleotide polymer in either single or double stranded form. The nucleic acid may include natural nucleotides or analogues of natural nucleotides that function in a similar manner. Nucleic acids may be derived from a variety of sources including, but not limited to, naturally occurring nucleic acids, clones, synthesis in solution or solid phase synthesis. A nucleic acid may refer to a polymeric form of nucleotides of any length, the nucleotides may be deoxyribonucleotides, ribonucleotides or peptide nucleic acids (PNAs) that comprise purine and pyrimidine bases or other natural or derivatised nucleotide bases. The sequence of the nucleotides may be interrupted by non-nucleotide components. Examples of nucleic acids which may be used in the invention include genomic DNA, RNA, cDNA and cRNA. Preferably the nucleic acids are labelled before use. The label may be a chemiluminescent, fluorescent or radioactive label.

Genomic DNA may be the DNA which comprises the genome of a cell or organism. Genomic DNA may include chromosomal DNA. Genomic DNA may also include non-chromosomal DNA such as mitochondrial DNA. Genomic DNA used in the method of the invention may be fragmented. The genomic DNA used may represent all or just a part of the genome. The genomic DNA may be cloned genomic DNA, which may be in a vector such as a BAC, YAC, P1 clone or cosmid.

Oligonucleotides identified by the method of the invention may then be selected for use in further analysis of nucleic acids from the second species or distinct variety of the first species.

Preferably a microarray is used only once as it is difficult to ensure that all the nucleic acid previously applied has been removed. Preferably, one microarray is used with genomic DNA to identify the relevant probes/oligonucleotides, and a second identical microarray is used in subsequent studies using only the identified probes.

The aim of the present invention is to identify which of the oligonucleotides/probes on a microarray derived from a first species can be used to analyse the corresponding nucleotide sequence from a second species or a distinct variety of the first species. By identifying and selecting the subset of oligonucleotides/probes on the microarray which hybridise to the genomic DNA of the second species, or distinct variety of the first species, the microarray can be tailored for use with the second species, or distinct variety of the first species. Having identified the subset of oligonucleotides/probes on the microarray which hybridise to the genomic DNA of the second species, or distinct variety of the first species, a mask may be generated which is specific to those oligonucleotides/probes. The mask may then be used when the microarray is used to analyse further the second species, or distinct variety of the first species, to ensure that only the selected probes are included/considered in the analysis. Probes which do not hybridise to the genomic DNA of the second species, or distinct variety of the first species, are therefore not considered in further analysis and thus cannot dilute or influence this analysis.

The use of genomic DNA from the second species, or distinct variety of the first species, to select the probes/oligonucleotides on the microarray ensures that as far as possible all the genes in the genome of the second species, or distinct variety of the first species, which are present on the array are represented in the probes identified and selected for use in further analysis. By using the genomic DNA to select the probes, the

selection is not biased to only the genes expressed in the sample material as all genes are represented in the genomic DNA. Preferably, the first and second species are related. Preferably, the first and second species are species of plant, animal, fungi or protist.

Preferably, the method of the invention can be used to identify and select oligonucleotides on a microarray which can be used to analyse gene expression and/or gene transcripts in a second species or a distinct variety of the first species. Gene transcripts include, but are not limited to, pre-mRNA transcripts, transcript processing intermediates, mature mRNA ready for translation and transcripts of the gene or genes, or nucleic acids derived from the mRNA, such as cDNA and cRNA.

The identified/selected oligonucleotides may be used to study the pattern of gene expression in the second species or distinct variety of the first species. For example, to analyse differences in gene expression between different tissues, at different times of development or in response to different environments or conditions.

The oligonucleotides identified by the method of the invention may be used to analyse changes within the relative RNA transcription. Alternatively, or additionally, the selected oligonucleotides may be used to analyse changes or differences between cross-species or cross-variety DNA, such changes may include deletions; chromosome rearrangements or insertions; polymorphisms such as single nucleotide polymorphisms (SNPs); and RNA changes other than simple gene transcripts. For example, the method according to the invention may be used to identify nucleotide sequences that have been deleted from the second species or distinct variety of the first species.

The method of the invention may be used to identify and select oligonucleotides/probes which can be used to analyse the transcriptome of the second species, or the distinct variety of the first species. Wherein the transcriptome may be the full complement of activated genes, mRNA, transcripts or other forms of RNA in a particular tissue at a particular time in the second species, or the distinct variety of the first species.

The oligonucleotides identified by the method of the invention may also be used to produce genetic maps and identify biallelic markers.

The invention may also be used to identify and select probes which may be used to identify a series of homologous genes to species one on a large cloned fragment of genome from species two using a vector associated construct preferably but not restricted to Bacterial Artificial Chromosomes (BACs) or Yeast Artificial Chromosomes (YACs). In one embodiment the cross species hybridisation would identify potential homologues between the species two BAC or BACs and annotated genes from species one which would then be analysed for observed co-linearity between the genomes of the species. Within this embodiment species one would most profitably be a species with superior genome annotation and the purpose of the analysis may be to generate information or markers for genome mapping or marker assisted breeding by comparative genomics.

The oligonucleotides/probes selected by the method of the invention may be used to compare two or more varieties of the second species, or two or more distinct varieties of the first species, and may allow differences between the species or varieties to be identified.

Oligonucleotides/probes selected or identified in the method of the invention may be used as primers for amplification, such as by PCR, or

for sequencing. The identified probes may be used as primers for the validation of genome information derived from probe sets.

A microarray is well known in the art, and the skilled man would understand the term. By way of background only, a microarray for use in the invention typically comprises a number of probe sets, each probe set being specific to a gene transcript from the species from which the array is derived. Each probe set may comprise between about 11 and about 20 probes which bind at various positions on the same gene transcript. The probes may be included as probe pairs, in which each probe pair may comprise a perfect match (PM) and a mismatch (MM) oligonucleotide probe. Preferably in the method of the invention only PM probes are considered to be informative.

Preferably the oligonucleotides/probes are from about 15 to about 80 nucleotides in length. More preferably the oligonucleotides/probes are from about 20 to about 30 nucleotides in length. More preferably the oligonucleotides/probes are about 25 nucleotides in length. If included, the mismatch (MM) probes preferably have a mismatch base when compared to the gene in the first species. Preferably the mismatch is around the middle of the probe. Preferably if a mismatch probe is used in the method of the invention it is considered as an oligonucleotide of very similar binding chemistry to the perfect match probe and thus can provide a measure of non-specific binding otherwise known as 'background'.

Usually when a microarray is analysed for hybridisation to applied nucleic acid each probe set relating to an individual gene is analysed as a single unit and a binding/hybridisation score is given to the probe set as a whole. The method of the present invention allows full probe sets and individual probes within a probe set to be selected for, or removed from, analysis when the microarray is used to analyse the corresponding

nucleotide sequence in the second species, or in a distinct variety of the first species. By excluding probes in a probe set which do not bind to the genomic DNA of the second species, or a distinct variety of the first species, the probe set can be tailored to a different species or distinct variety and thus become more discriminating and useful. If mismatch probes have been included on the microarray the user may chose to exclude them when tailoring the microarray. Polymorphisms, even at a single nucleotide, between the first species from which the microarray is derived and the second species, or in a distinct variety of the first species, may result in one or more of the probes in a probe set not binding/hybridising to the genomic DNA, if these probes are not removed from further analysis the probe set will give a lower binding/hybridisation score than if the probes which do not bind the genomic DNA were removed. If however those members of the probe set which do not hybridise sufficiently to the genomic DNA of the second species are removed from analysis the remaining members of the probe set can be used in the subsequent analysis of the corresponding nucleotide sequences in the second species, or in a distinct variety of the first species. Preferably for a probe set to be selected by the method of the invention the derived set (that is, the set produced when only those probes which hybridise to the added genomic DNA are selected) must include at least one probe. Ideally the derived probe set should include multiple discriminating probes/oligonucleotides, preferably at least two probes, more preferable at least three probes. Preferably the selected probes are perfect match probes.

Once the probes which hybridise to the genomic DNA of the second species, or a distinct variety of the first species, have been identified/selected a mask may be generated defining only those probes which are to be retained, or used in further analysis of the second species, or a distinct variety of the first species. By applying the mask to

a microarray of the first species it is effectively converted into a microarray directed to the second species, or the distinct variety of the first species.

Hybridisation of genomic DNA to the oligonucleotides/probes on the microarray may be determined by using genomic DNA which has been labelled, for example with a fluorescent, chemiluminescent or radioactive label, and then screening the microarray for the label to identify oligonucleotides or probes in the mircorarray to which the genomic DNA has hybridised. Methods to isolate genomic DNA are well known to those skilled in the art, as are methods to label the DNA. For example the Bioprime® DNA labelling system from Invitrogen may be used to label genomic DNA.

By adjusting the stringency of the hybridisation conditions, or the subsequent washing conditions, the pattern of hybridisation of the labelled genomic DNA to the microarray can be varied. Different oligonucleotides will be selected depending on the conditions used. For example, under conditions of high stringency only genomic DNA identical to the probes will hybridise, whereas under a lower stringency some degree of mismatch will be allowed and hence more probes are likely to hybridise to the genomic DNA. Different masks may be generated to reflect the different conditions. The intention of adjusting the stringency of the hybridisation and/or wash conditions is to identify the optimum conditions when the genomic DNA hybridises to probes from as many genes as possible with minimal non-specific hybridisation.

The person skilled in the art will appreciate how to adjust and optimise the stringency conditions for hybridisation and washing in nucleic acid hybridisation/binding studies.

The probe set members which are excluded from use in further analysis of the second species, or the distinct variety of the first species, may themselves provide information about the differences between the first and second species, or distinct variety of the first species, in terms of a single nucleotide polymorphism (SNP), a deletion or mapping studies including the comparative genomics of cross-species colinearity.

GeneChips® may be generated to whole genome sequence orthan restricted to transcribed regions. One such type of whole genome chip is commonly known as a 'tiling chip'. Tiling chips could also be used with the method of the invention to investigate control elements, SNPs, deletions and cross-species colinearity of the genome between species or between distinct varieties of the same species. Such chips, may also be used to identify and annotate potential new coding regions via in-silico prediction. For all GeneChips® the initial labelling of the genomic DNA to allow selection of the useful probes and the generation of a mask production should preferably not be subject to asymmetric amplification in order that an assessment of genome duplication and gene family number may also be attempted.

The microarray used may be commercially prepared, such as those made by Affymetrix® or Nimblegen®.

The method of the invention may be adapted to use more than one microarray. Two or more microarrays from different species may be used in parrallel to allow the identification and selection of oligonucleotides on each of the microarrays which may be then be used to analyse the corresponding nucleotide sequences in a species or a distinct variety different to that used to produce the microarrays. For example, a tomato microarray and an Arabidopsis microarray may be used to study a

primula. Each microarray may contribute probes representative of different genes to the study of the primula.

According to a further aspect the invention provides a method of analysing nucleic acids in a second species, or a distinct variety of a first species using a microarray from a first species, comprising:

applying genomic DNA of the second species, or a distinct variety of the first species, to the microarray derived from a first species;

identifying probes/oligonucleotides on the microarray to which the genomic DNA has hybridised;

selecting the probes/oligonucleotides on the microarray to which the genomic DNA has hybridised for further analysis;

applying mRNA, cDNA or cRNA from a tissue of the second species, or distinct variety of the first species, to a microarray derived from the first species;

analysing the pattern of hybridisation of the mRNA, cDNA or cRNA to the selected probes/oligonucleotides.

This method may be used to study gene expression in a second species, or a distinct variety of a first species using a microarray from a first species.

Preferably the genomic DNA, mRNA, cDNA and/or cRNA is labelled before use. The label may be a fluorescent, chemiluminescent or radioactive label.

By analysing the pattern of hybridisation of the mRNA, cDNA or cRNA to the selected probes the genes which are expressed, and those which are not expressed, in the tissue form the second species, or distinct variety of the first species, can be determined. This method allows the transcriptome of a second species, or distinct variety of the first species, to be studied using a microarray derived from a first species.

According to another aspect the invention provides the use of the probes identified according to the first or second method of the invention to study gene expression in a second species, or a distinct variety of the first species.

The selected probes may also be used to study change in gene structure between a first species and a second species, or a distinct variety of the first species. The changes may include deletions, insertions or mutations.

According to a further aspect the invention provides a kit for selecting oligonucleotides on a microarray comprising a microarray derived from a first species and instructions to use the method of the invention with genomic DNA of a second species or a distinct variety of the first species.

According to a further aspect the invention provides a kit for analysing gene expression in a second species or a distinct variety of a first species comprising a microarray derived from a first species and instructions to use the microarray according to the method of the invention.

An advantage of the method of the invention is that a microarray already available can be tailored for use with a species or a variety for which a microarray is not available. For example, microarrays are currently only available for a small number of plant and animal species. The cost of producing a microarray for a specific species can cost tens or hundreds of thousands of pounds, and can take several years. Preferably the invention is used where the first and second species are related. In one embodiment the methods of the invention may be used on Arabidopsis arrays for Brassica crops, for example oilseed rape, broccoli, cabbage, cauliflower, Brussels sprout, kale, Chinese cabbage etc, or to other related species.

The present invention provides a tailored approach for cross-species/variety hybridisation of crop species nucleic acid to high density oligonucleotide arrays designed for model or extensively sequenced organisms, such as Arabidopsis. For the purposes of the present invention cross-species hybridisation is taken to mean the hybridisation of nucleic acid fragments of one species with nucleic acid fragments of a related species or distinct variety of the same species.

The procedure for the hybridisation of genomic DNA to the microarray is as for mRNA and is described in DE69625920T. After staining the microarray the microarray is scanned to identify where the genomic DNA has bound, as described in DE69625920T. The software used to analyse the microarray is described in detail in US5547839, US5578832 and US5631734. The software generates a hybridisation intensity file (CEL) containing the statistics of the array eg the 75th percentile of intensities, standard deviation of pixel intensities and probe co-ordinates which represent the physical location of the probes on the array. To further analyse the data to generate probe masks, which determine which probes should be considered in subsequent analysis, the computer language PERL (see e.g., Wall Christiansen and Orwant, Programming Perl, 3rd Ed, O'Reilly and Associates (2000)) is typically used to construct the necessary computer programs. Although equivalent scripts and programs can readily be developed in other computing languages by a person skilled in the art (including, but not limited to, C++, Java, Visual Basic).

First a perl script is constructed to extract probe co-ordinates with a hybridisation intensity above background. The term background refers to a calculated estimate of the intensity level that represents non-specific hybridisation or other interaction between the hybridising target and components of the array. Fluorescence of the array components may also

contribute to background. Background may be calculated as the mean intensity of negative control probes. The negative control probes may be Bio B, Bio C, Bio D, Cre etc, or other oligonucleotide probes selected from species or varieties other than the 'reference species' organism or the cross species organism.

The oligonucleotide probe co-ordinates selected by the perl script may include both perfect match and mismatch probes. In a cross-species genomic DNA hybridisation it is conceivable, albeit unlikely, that some of the mismatch probes will hybridise more efficiently to the cross-species target DNA sequence.

A second perl script may be developed to eliminate mismatch probes with hybridisation intensity above background and with higher hybridisation intensity than perfect match probes. To achieve this the perl script may use as an input, a file consisting of only perfect match co-ordinates and the output file of the first perl script. The output file generated may consist of only perfect match probe co-ordinates with hybridisation intensities greater than the estimated background for the genomic DNA hybridisation. These perfect match oligonucleotide probes, which share high sequence similarity with the cross-species target, represent probes selected for the analysis of cross-species transcripts hybridised to the 'reference species' array.

The selection of the corresponding mismatch oligonucleotide probes may be carried out using a third perl script. A third perl script may be constructed to complete the process of generating a chip description file (CDF) for the cross-species organism. This perl script may use as an input, the output (selected perfect match probes) of the previous script and the chip description file of the 'reference species' organism. The X co-ordinate for both perfect match and mismatch probes for each gene

sequence on the array is identical. This information is coded into the perl script to enable the selection of both perfect match and mismatch probes from the 'reference species' organism's CDF to construct a new chip description file for the cross-species. The new CDF forms the mask for studying the cross species. The probes excluded from CDF construction may be used to construct a probe sensitivity index (PSI) file. The PSI file may then be used to train a large data set in order to ensure consistency of data stored in a database.

All the software used in the computation of gene expression levels requires a hybridisation intensity file (CEL) of the type described above and a chip description file (CDF) for the array type carrying the hybridised target transcripts. The chip description file is a library file consisting of gene (probe set) IDs, the corresponding co-ordinates of their probe sequences on the array and other software parameters.

The first species or 'reference species' CDF has normally been derived by a commercial vendor such as Affymetrix™ and is made available to the general public from their website and on media distributed with their GeneChips™ for use in analysing their proprietary material.

In a further aspect of the invention, a BLAST (Alschul et al., J. Mol. Biol. 215;403-410 (1990) output file may be generated in silico by comparing cross-species/second species/distinct variant nucleic acid sequence in a database to nucleic acid sequence represented on the oligonucleotide array. This output file may be parsed with another perl script to identify oligonucleotide probes with 100% sequence identity to the cross-species sequence. The probe selected may then be used to construct a chip description file for the cross-species organism as described above.

The invention allows the informed selection of probe pairs (putative or actual match and mismatch) to allow high throughput genome-wide screening of the expression patterns of genes from genomes of species related to but not identical to the genome of at least one reference extensively-sequenced species such as, but not limited to, Arabidopsis, tomato, rice, mouse, Human, C. elegans, Bacillus sp., Drosophila, Chimpanzee, Chicken and the SARS virus.

According to another aspect the invention provides a computer system for selecting oligonucleotide probes comprising:

a co-ordinate extraction means arranged to extract the co-ordinates of probes on a microarray derived from a first species to which genomic DNA from a second species, or a distinct variety of the first species, has been applied which display a hybridisation intensity with the genomic DNA that is above background to generate a match co-ordinate output;

a mismatch elimination means arranged to identify and eliminate mismatch probes with a higher hybridisation intensity than perfect match probes from the match co-ordinate output to generate a perfect match co-ordinate output;

a chip description file (CDF) generation means arranged to compare the first species CDF with the perfect match co-ordinate output and to generate a further CDF comprising the co-ordinates present in both the first species CDF and the perfect match output.

The further CDF may be used as a mask in further analysis of the second species, or distinct variety of the first species, which determines/selects which probes/oligonucleotides on the microarray of the first species are to be considered.

The computer system may also comprise a background determination means.

According to a further aspect the invention provides a computer system for generating a mask comprising:
a reader arranged to detect where genomic DNA has hybridised to a probe on a microarray and to produce data indicative of where hybridisation has occurred; and
a processor arranged to combine the data from the reader with a CDF for the microarray to produce a mask.

Preferably the genomic DNA hybridised to the microarray probes is from a species or variety different to that used to make the microarray.

Preferably the data generated by the reader is a set of co-ordinates corresponding to the probes which hybridised to the genomic DNA.

Preferably the mask is a computer programme arranged to operate a reader so that when the mask is applied the reader only considers specific coordinates on a microarray which correspond to oligonucleotides/probes which hybridised to the genomic DNA. The mask may alternatively be defined as a further CDF.

The mask may be used to tailor a microarray from a first species to a different species, or distinct variant of the first species. By considering only those probes which hybridised to the genomic DNA, subsequent analysis is not diluted by the inclusion of probes that will not bind to DNA of the second species or the distinct variant of the first species.

Preferably the reader is a device with the capacity to analyse all the probes on a microarray. Each probe on the microarray has unique coordinates. Preferably any nucleic acid, for example genomic DNA, mRNA, cDNA or cRNA, added to the array is labelled so that the reader can detect where hybridisation to a probe on the array has occurred.

According to a further aspect the invention provides a method of making a mask comprising the steps of:
applying genomic DNA from a second species or a distinct variety of a first species to a microarray derived from a first species;
analysing with a reader the microarray to determine which probes on the array have hybridised to the genomic DNA;
comparing the CDF file for the microarray with the data from the reader;
generating a mask which represents the coordinates of probes on the microarray which hybridised with the genomic DNA.

Preferably at least one of the steps of making a mask is undertaken on a computer.

According to another aspect the invention provides a data carrier carrying data arranged to control a computer system to carry out the method of making a mask according to the invention or to operate as a computer system according to the invention.

The skilled man will appreciate that preferred aspects of the invention discussed with reference to only one aspect of the invention can be applied to all aspects of the invention.

Preferred embodiments of the invention will now be described merely by way of example with reference to the accompanying drawings, in which:

Figures 1a and 1b - illustrates the effect of increasing hybridisation stringency on the number of probes on a mircoarray which bind to the applied genomic DNA. More specifically, Figures 1a and 1b illustrate the number of *Arabidopsis thaliana* perfect match (PM) probes and probe sets from the ATH1-121501 GeneChip® array used to study the transcriptome of *Brassica oleracea* var. *alboglabra cv.* A12DHd at different levels of DNA hybridisation. Figure 1a shows the relative distribution of PM probes and Figure 1b illustrates the absolute number of PM probes and probe sets, as a function of the DNA hybridisation intensity thresholds used to generate the probe mask files. A threshold criterion of two PM probes was set for probe set inclusion. In Figure 1a the number of probes selected per probe set are indicated directly on the diagram by arrows. In Figure 1b, filled circles are scaled to the left-hand y-axis (i.e. probe sets used in probe mask files) and unfilled circles are scaled to the right-hand y-axis (i.e. perfect match probes used in probe mask files). The data was obtained by hybridising genomic DNA from *B. oleracea* to the *A. thaliana* ATH1-121501 GeneChip® microarray.

Figures 2a – 2f– illustrate probe set signals of genes in control (P-replete) *Brassica oleracea* var. *alboglabra cv.* A12DHd estimated following probe selection compared to probe set signals estimated without probe selection. Data are presented using probe mask files generated at 50, 100, 200 and 400 respectively (Figure 2a – 2d). Mean values Figure 2e and ranked coefficient of variation Figure 2f of probe set signals of control (P-replete) *B. oleracea* as a function of the DNA hybridisation intensity threshold was used to generate probe mask files for the transcriptome analysis. In Figure 2f the DNA hybridisation intensity threshold used to generate probe mask files is indicated by 13 different intensity

lines from left to right as indicated representing DNA hybridisation intensity thresholds of : 0, 50, 100, 150, 200, 300, 400, 500, 600, 700, 800, 900, 1000. In all panels, total RNA samples were extracted from the shoots of hydroponically-grown control (P-replete) B. oleracea (n = 4).

Figures 3a - d – depict 'Volcano' plots illustrating the fold-changes (i.e. twice the log2 of the ratio of means for each gene) and inverse significance (i.e. log10 of the reciprocal of the P-value derived from a one-sample, two-tailed, Student's t-test) in gene expression differences between control and P-starved *Brassica oleracea* var. *alboglabra cv.* A12DHd. Total RNA samples were extracted from control *B. oleracea* shoots and from the shoots of plants grown in the absence of P for 100 h (n = 4). (a) no probe selection used during transcriptome analysis, (b), (c), (d) using probe mask files during transcriptome analysis, generated at DNA hybridisation intensity thresholds of 200, 300 and 1000 respectively.

Figures 4a - d – illustrate gene regulation under P-starvation in *Brassica oleracea* var. *alboglabra cv.* A12DHd as a function of the DNA hybridisation intensity threshold used to generate probe mask files for the transcriptome analysis. Total RNA samples were extracted from control *B. oleracea* shoots and from the shoots of plants grown in the absence of P for 100 h (n = 4). (a), (b) genes significantly regulated under P starvation at $P < 0.05$ and $P < 0.01$ respectively. (c) genes regulated $\pm$ > 1.5-fold under P starvation. (d) genes significantly regulated $\pm$ > 1.5-fold under P starvation ($P < 0.05$).

Figures 5a and 5b - shows dChip (Li and Wong, Proc. Natl. Acad. Sci. USA, 98; 31-36 (2001)) software data of probe response patterns for the probe set, 246745_at. The upper line of the graphs (labelled PM) represents perfect match probe intensities and the lower line (labelled MM) represents mismatch probe intensities. Figure 5a depicts the probe response pattern for 246745_at on an array (ATH1 – 12501) hybridized (see Patent No DE69625920T for all hybridization conditions) with an Arabidopsis (B1798_rep1) cRNA target. Figure 5b represents a probe response pattern for 246745_at on the same array type hybridized with Brassica (Bo + P_repl) cRNA target.

Figure 6 - illustrates gene expression levels computed with GeneChip® software for P treated Arabidopsis and Brassica using an Arabidopsis microarray. The data shows that when a CDF designed for Brassica is used with an Arabidposis microarray an increased signal is observed.

Figure 7 - illustrates schematically a computer system according to the invention.

DNA hybridisation and probe selection

Arabidopsis thaliana ATH1-121501 GeneChip® microarrays (available from Affymetrix™) were used to study the transcriptome of *B. oleracea*. Although DNA sequences are highly conserved between *A. thaliana* and *B. oleracea* (Cavell et al., 1998, Genome 41, 62-69; O'Neill and Bancroft, 2000, Plant J. 23, 233-243), sequence polymorphisms between the two species are likely to result in an underestimate of transcript abundance if all probes are used within individual probe sets. Therefore, we selected subsets of PM probes from each probe set on the *A. thaliana*

GeneChip® microarray based on the hybridisation efficiency of genomic DNA from *B. oleracea* to homologous *A. thaliana* GeneChip® microarray probes/oligonucleotides. Prior to the selection of probes by genomic DNA hybridisation, it was confirmed that ATH1-121501 GeneChip® probes were likely to hybridise efficiently to genomic DNA targets from *B. oleracea*, using publicly available *B. oleracea* genomic shotgun sequence data. As expected, there was substantial sequence homology between these two species of *Brassicaceae* (e.g. see http://atensembl.arabidopsis.info/).

Genomic DNA from *B. oleracea* was biotin-labelled and hybridised to the *A. thaliana* ATH1-121501 GeneChip® microarray. Probe sets were selected for subsequent transcriptome analyses if the probe set was represented by at least two PM probes with hybridisation intensities above a set threshold. Probe sets were selected using text-extraction algorithms written in the Perl programming language. The method was optimised empirically by generating 13 probe mask files with DNA hybridisation intensity thresholds ranging from 0 (i.e. no probe selection) to 1000. The different thresholds reflect different stringencies of hybridisation and washing conditions. The aim is to identify the optimal conditions when as may genes as possible are represented in the selected probes but there is minimal non-specific binding. These 13 probe mask files were subsequently used in turn to quantify the transcriptional response of *B. oleracea* to a mineral nutrient (phosphorus; P) stress. In practice the aim would be to identify the optimal mask and use this in future stuides.

Arabidopsis thaliana PM probes hybridised to the *B. oleracea* genomic DNA as represented in Figure 1. When the DNA hybridisation intensity threshold was increased from 0 to 1000 during probe mask file generation, that is the binding of genomic DNA to the probes on the microarray decreased. PM probe retention in the probe mask files

decreased rapidly (Figure 1b). However, the retention of probe sets was less sensitive to increases in DNA hybridisation intensities during probe mask file generation, since a minimum of two PM probes were required to retain a probe set (Figure 1b). For example, although the probe mask file generated at a DNA hybridisation intensity threshold of 300 masked 56 % of all PM probes, only 3.9 % of available *A. thaliana* probe sets were lost.

Testing DNA probe selection: transcript abundance in control *Brassica oleracea*

Brassica oleracea were grown hydroponically using physiological techniques described elsewhere (Hammond et al., 2003 Plant Physiol. 132, 578-596). Control plants were supplied with a full nutrient solution throughout the experiment; treated (P-starved) plants were supplied with a nutrient solution containing no P for 100 h. Transcriptional responses of *B. oleracea* to phosphate stress were determined by challenging *A. thaliana* ATH1-121501 GeneChip® microarrays with total RNA extracted from control and treated plants (n = 4). Following scanning of the GeneChip® microarrays, raw cell intensity data files (.cel files) were loaded into GeneSpring (Silicon Genetics, CA, USA) using 13 probe mask files. For each of the 13 probe selection conditions, data were prenormalised using Robust Multichip Average (RMA) algorithms (Irizarry et al., 2003 Biostatistics 4, 249-264). Subsequently, data from the eight GeneChip® microarrays (four control and four P-starved samples) were treated as 13 individual 'experimental scenarios' to evaluate probe selection stringency. Within each 'scenario', data were further normalised within GeneSpring; the signal value from each replicate P-starved plant was normalised to its corresponding control sample.

The use of probe masks increased estimates of transcript abundance in *B. oleracea* in both control and P-starved samples (data shown for control samples; Figure 2a-e). The probe mask file generated with a DNA hybridisation intensity threshold of 50 did not affect estimates of transcript abundance; only 123 of 250,195 available PM probes were excluded from this probe mask file. Thus, all probe sets were retained for subsequent transcriptome analysis (Figures 1b, 2a). However, probe mask files generated with DNA hybridisation intensity thresholds of 100 and above increased estimates of transcript abundance compared to using no probe selection (Figure 2b-d). Estimates of transcript abundance in *B. oleracea* increased by up to three-fold when probe masks were used (Figure 2e). Increasing the DNA hybridisation intensity threshold during probe mask file generation also increased the coefficient of variation (CV) calculated for ranked transcripts (probe sets) across the four replicate control arrays (Figure 2f). Thus, if stringent probe selection criteria are to be used to analyse experimental transcriptome data, it may be appropriate to increase biological replication. However, the CV of control arrays were minimally affected using probe mask files generated at DNA hybridisation intensity thresholds below 500 (Figure 2f).

Testing DNA probe selection: gene regulation under P stress in *Brassica oleracea*

Gene expression was normally distributed in control and P-starved samples when experiments were analysed using any of the 13 probe mask files (data not shown). For each experimental interpretation, fold-change differences in gene expression were calculated as twice the log2 of the ratio of means for each gene in the treated relative to the control sample. One-sample, two-tailed, Student's t-tests were used to test if transcript abundance was significantly different between control and P-starved samples. Increasing the DNA hybridisation intensity threshold from 0 to

500 for probe mask file generation increased the sensitivity of the ATH1-121501 GeneChip® microarray to detect regulation of gene expression in *B. oleracea* under P stress, both in terms of statistical significance and in terms of fold-differences in gene expression between P-starved and control samples (Figures 3, 4).

Estimates of fold-differences in gene expression (P-starved samples versus control samples) increased using probe mask files generated with DNA hybridisation intensity thresholds up to 500 (Figure 3, 4c). 'Volcano' plots summarise these effects (Figure 3a-c). In volcano plots, the y-axis represents the log10 of the reciprocal of the t-test P-value; the x-axis represents the fold-change in gene expression. There was a five-fold increase in the number of genes identified as differentially regulated >1.5-fold under P stress using a probe mask file generated with a DNA hybridisation intensity threshold of 500 (Figures 3, 4c). The use of probe mask files generated with higher DNA hybridisation intensity thresholds (up to 1000) resulted in a substantial loss of probe sets available for transcriptome analysis (Figures 1b, 3d). However, the number of genes identified as differentially regulated >1.5-fold was still greater when using the probe mask file generated with a DNA hybridisation intensity threshold of 1000 than when probe selection was not used, even though only 4.3 % of the total number of PM probes on the GeneChip® microarray were used (Figure 4c). The perturbed boundary conditions of volcano plots are still clearly visible under conditions of high-stringency probe selection compared to when no probe selection was used (Figure 3d).

Estimates of the number of genes significantly differentially regulated under P stress (at $P < 0.05$) also increased when probe mask files were used (Figure 4a). For example, using a probe mask file generated with a DNA hybridisation intensity threshold of 150, 1,610 genes were estimated

as significantly differentially regulated, compared to 1,514 genes regulated using no probe selection. At a more stringent significance threshold ($P < 0.01$), the number of genes estimated as significantly differentially regulated under P stress also increased with the use of probe mask files. For example, using a probe mask file generated at a DNA hybridisation intensity threshold of 200, 328 genes were estimated to be significantly differentially regulated, compared to 286 genes using no probe selection. Probe mask files generated using DNA hybridisation intensity thresholds above 400 resulted in a reduced ability to detect transcriptional regulation under P stress in *B. oleracea*, due to the loss of available probe sets for the analysis.

An optimal probe selection strategy for transcriptional analysis of *B. oleracea* is to use a probe mask file generated at a DNA hybridisation intensity threshold of 300. Although estimates of fold-differences in gene expression in P-starved versus control samples were greatest using probe mask files generated with DNA hybridisation intensity thresholds of 400 and 500 (Figure 4c, d), there was a significant loss of available probe sets for transcriptome analysis at DNA hybridisation intensity thresholds > 300. Further, estimates of the number of significantly regulated genes declined using probe mask files generated at DNA hybridisation intensity thresholds > 300.

Biological significance of genes regulated under P stress in *Brassica oleracea*

To interpret the biological significance of the P stress response in *B. oleracea*, experimental data were analysed using a probe mask file generated at a DNA hybridisation intensity threshold of 300. In total, 1,559 genes were significantly differentially regulated in the shoots of *B. oleracea* following the withdrawal of P from the nutrient solution ($P < 0.05$; Supplementary Table I. Of these genes, 980 had higher transcript abundance in P-starved samples than in control samples and 579 had lower transcript abundance in P-starved samples than in control samples.

Sequence polymorphisms between *A. thaliana* and *B. oleracea* are likely to result in the identification of homologous genes or alternative members of gene families in addition to putative gene orthologues. For this reason, previously published studies on the response of *A. thaliana* to P starvation (Hammond et al., 2003, Plant Physiol. 132, 578-596; Hammond et al., 2004 Ann. Bot. 94, 323-332; Wu et al., 2003 Plant Physiol. 132, 1260-1271) at the level of the individual gene and also at the level of the gene family and functional category were compared with the results above. Several homologous genes responded similarly to P starvation in *B. oleracea* and *A. thaliana*. For example, SQD2 (At5g01220) and MGD2 (At5g20410), which are involved in sulpholipid and galactolipid biosynthesis, increase their expression in responses to P starvation in *A. thaliana* (Essigmann et al., 1998 Proc. Natl Acad. Sci. USA 95, 1950-1955; Yu et al., 2002 Proc. Natl Acad. Sci. USA 99, 5732-5737; Hammond et al., 2003, Plant Physiol. 132, 578-596; Hammond et al., 2004 Ann. Bot. 94, 323-332; Wu et al., 2003 Plant Physiol. 132, 1260-1271; Kobayashi et al., 2004 Plant Physiol. 134, 640-648). Homologues of these genes had higher hybridisation signals in P-starved samples than

in control samples. The signal values for SQD2 and MGD2, obtained using the probe mask file generated with a DNA hybridisation intensity threshold of 300, were based on six and four probes respectively in contrast to 11 probes when no probe selection was used. The use of probe selection increased hybridisation signals in P-starved samples compared to control samples to a greater amount (137.50 ± 33.06, SQD2; 52.97 ± 2.94, MGD2; average signal ± S.E.M.) than in the absence of probe selection (71.77 ± 17.17, SQD2; 13.84 ± 0.51, MGD2).

Ribonucleases, phosphatases and phosphodiesterases are involved in the recycling of P in plants during P starvation (Bariola et al., 1994 Plant J. 6, 673-685; van der Rest et al., 2003 Plant Physiol. 130, 244-255; Hammond et al., 2003, Plant Physiol. 132, 578-596; Hammond et al., 2004 Ann. Bot. 94, 323-332). The hybridisation intensity of the ribonuclease RNS2 (At2g39780) was higher in P-starved samples compared to control samples, and it may be involved in the release of P from internal sources during P starvation. The signal value for RNS2, obtained using the probe mask file generated with a DNA hybridisation intensity threshold of 300, was based on five probes in contrast to 11 probes when no probe selection was used. The use of probe selection increased hybridisation signals in P-starved samples compared to control samples to a greater amount (882.22 ± 49.89, RNS2; average signal ± S.E.M.) than in the absence of probe selection (676.38 ± 20.51, RNS2). Higher hybridisation signals in P-starved samples compared to control samples were also observed for several other genes that belong to gene families or functional groups that have previously been observed responding to P starvation in *A. thaliana* (Hammond et al., 2003, Plant Physiol. 132, 578-596; Hammond et al., 2004 Ann. Bot. 94, 323-332; Wu et al., 2003 Plant Physiol. 132, 1260-1271). These include several genes involved in the transport of phosphate and phosphate containing compounds and carbohydrates, indicating a change in the internal

economy of P use in response to P starvation. Significantly higher hybridisation intensities for a sucrose synthase, a malic enzyme and a malate dehydrogenase in P-starved compared to control samples, and a lower hybridisation intensity for a pyruvate kinase gene in P-starved compared to control samples indicate an increase in alternative glycolosis reactions that conserve P during P starvation conditions (Hammond et al., 2004 Ann. Bot. 94, 323-332). Interestingly, the expression of several genes involved in photosynthesis had significantly higher hybridisation intensities P-starved compared to control samples. There are other notable similarities in the P starvation response of *B. oleracea* with published responses of *A. thaliana* to P starvation at the level of the gene family. These include transcription factors from the zinc finger, F-box, bHLH and myb families, genes involved in flavanoid biosynthesis and genes encoding proteins from the cytochrome P450, glycosyl hydrolase, and peroxidase families.

To summarise, the above data provides a novel probe selection method to study the transcriptome of a species for which GeneChip® microarrays are not available. Genomic DNA from *B. oleracea* was labelled and hybridised to the *A. thaliana* ATH1-121501 GeneChip® array. Perfect match *A. thaliana* probes which hybridised to the *B. oleracea* genomic DNA above selected hybridisation intensities were selected for subsequent *B. oleracea* transcriptome analysis using probe mask files generated using text-extraction algorithms.

Probe selection was tested by quantifying the transcriptional response of *B. oleracea* to a mineral nutrient (phosphorus; P) stress using probe mask files generated at DNA hybridisation intensity thresholds ranging from 0 (i.e. no probe selection) to 1000. Increasing the DNA hybridisation intensity thresholds for probe mask file generation increased the sensitivity of the GeneChip® microarray to detect regulation of gene

expression in *B. oleracea* under P stress. An optimum probe mask file was generated using a DNA hybridisation intensity threshold of 300; although 56 % of the *A. thaliana* PM probes were excluded from the subsequent transcriptome analysis, 96.1 % of the total available probe sets were retained. Up to 1,559 genes were significantly regulated in the shoots of *B. oleracea* under P stress ($P < 0.05$), including genes involved in signal transduction, flavanoid biosynthesis and the recycling of P within plants.

Figures 5a, 5b and 6 further support the need to select the probes within a probe set on a microarray to allow a microarray to be used more effectively with a species or variety different to that used to generate the microarray. Figure 5a and b show the hybridisation of Arabidopsis cRNA (Figure 5a) and Brassica (cRNA) to the probe set 246755_at on an Arabidopsis microarray. Each of Figure 5a and 5b includes 11 probe pairs. In Figure 5a all eleven pairs respond actively to the Arabidopsis target. However, in Figure 5b for the Brassica (cross-species) target only a subset of the 11 probes respond to the target. This is evident by the fact that for probe pairs 1-7 (Figure 5b) the PM and MM curves are indistinguishable. Ymax for the graphs represents the highest probe signal intensity. These unresponsive probes would be eliminated from the Arabidopsis (ATH1) chip description file (CDF) in the construction of the cross-species (Brassica) chip description file.

Figure 6 illustrates gene expression levels computed with GeneChip® software. There are three main columns; the first column (probe set ID) consists of gene (probe set) IDs. The second (ATHT_Bo+P) and third (Brassica_Bo+P) columns, which are subdivided into two columns, denoting the number of probe pairs (Stat Pairs) and expression levels (signal), represent a sample (Bo+P) analysed with the model organism CDF (ATH1 - *Arabidopsis thaliana*) and the cross-species CDF

(Brassica). In the Brassica CDF the number of probes in each probe set has been adjusted to reflect binding to the genomic Brassica DNA. The data in Figure 6 illustrates that by only using the best probes in the probe set then much better expression levels are observed.

The data illustrates that high level expression values are generated with the cross-species CDF. Since the mean of perfect match and mismatch probe pairs in a probe set is output by the software as the expression estimate of transcripts, probes in the ATH1 CDF which are not responsive to Brassica transcripts will lead to an attenuation of signal for the probe set, generating and inaccurate expression estimate of the manuscript. In a typical experiment the background signal is usually less than 100. Therefore transcripts with expression levels below background are usually identified as undetectable in the sample being interrogated.

Figure 7 shows a computer system which is suitable for carrying out an embodiment of the invention. The system is intended to be accessed by a user of a personal computer system, or terminal means, such as the system shown in Figure 7. The personal computer 100 comprises a display 102, processing circuitry 104, a keyboard 106, a mouse 108 and a reader 110. The processing circuitry 104 comprises a display driver 111, a processing unit 112, a hard drive 114 and a memory 116, all of which can communicate 120 with each other and with the reader 110.

The reader 110 is arranged to interrogate a microarray to determine which of the probes have hybridised to added nucleic acid. The nucleic acid may be genomic DNA, mRNA, cRNA or cDNA. The nucleic acid is labelled before use, and it is the label which allows hybridisation to the microarray to be detected. The reader may be a scanner, such as an Affymetrix G2500A GeneArray scanner.

Information/data from the reader 110 is feed to the processing unit 112 where it can be interrogated and analysed. The information/data from the microarray obtained by the reader 110 is processed and then displayed on the computer display 102 or printed out using a printer (not shown). The processing unit 112 is capable of comparing the data from the reader 110 with the CDF for the microarray to produce a new CDF.

MATERIALS AND METHODS

DNA hybridisation and probe selection

Probes from the *A. thaliana* ATH1-121501 GeneChip® microarray (Affymetrix, Santa Clara, CA, USA) were selected for transcriptome analysis of *B. oleracea* using a DNA-based probe selection strategy. Total genomic DNA was extracted from 5 g of *B. oleracea* var. *alboglabra cv.* A12DHd leaf tissue using a DNeasy Plant mini kit (Qiagen Ltd, Crawley, UK). DNA was labelled using the Bioprime DNA labelling System (Invitrogen Life Sciences, Carlsbad, California, USA) and subsequently hybridised to Affymetrix ATH1-121501 GeneChip® microarrays for 16 h at 45 °C using standard Affymetrix hybridisation protocols. Subsequently, the GeneChip® microarray was scanned on an Affymetrix G2500A GeneArray scanner and a cell intensity file (.cel file) was generated using Microarray Analysis Suite (MAS Version 5.0; Affymetrix). This .cel file contained the DNA hybridisation intensities between *B. oleracea* genomic DNA fragments and all *A. thaliana* probes. Perfect match probes from the .cel file were selected for subsequent transcriptome analysis using a text-feature extraction script written in Perl programming language (www.perl.com). The Perl script was designed to create probe mask (.cdf) files compatible with a range of microarray analysis software packages. A probe set was selected only when it was represented by at least two PM probes per probe set (i.e. a minimum of

50 bp homologous probe sequence of *A. thaliana* was required for subsequent transcriptome analysis of *B. oleracea*). There was no *a priori* indication of a suitable DNA hybridisation intensity threshold for probe mask file generation. Thus, the algorithm was designed to allow a user-specified DNA hybridisation intensity threshold for probe mask file generation to be set. Files were therefore generated using a range of DNA hybridisation intensity thresholds (from 0 to 1000) using a series of different scripts.

Testing the DNA probe selection strategy: determining the transcriptional response of *B. oleracea* to phosphate stress

Brassica oleracea var. *alboglabra cv.* A12DHd were grown hydroponically from seed in a system described in Hammond et al., (2003). Control plants were supplied with nutrient solution containing all nutrients throughout the experiment whilst treated plants were supplied with a nutrient solution containing no P for 100 h. Replicate samples (each replicate comprising 8-10 individual plants) were harvested from control and treated plants, mid way through the photo-period, and snap frozen in liquid nitrogen. All plants had approximately 6-8 leaves at harvest. Three biological replicates and one technical replicate were used for transcriptome analysis.

Tissue samples, previously stored at -70°C, were placed in liquid nitrogen before grinding. To each sample, 1 ml of TRIzol reagent (Invitrogen Life Technologies) was added and total RNA was subsequently extracted as described previously (Hammond et al., 2003 Plant Physiol. 132, 578-596). Total RNA yield and purity were determined using an Agilent 2100 Bioanalyser (Agilent Technologies). Approximately 5 μg of total RNA was reverse transcribed at 42 °C for 1 h to generate first strand cDNA using 100 pmol oligo dT(24) primer

containing a 5'-T7 RNA polymerase promoter sequence, 50 mM Tris-HCl (pH 8.3), 75 mM KCl, 3 mM MgCl2, 10 mM dithiothreitol (DTT), 10mM dNTPs and 200 units SuperScript II reverse transcriptase (Invitrogen Life Technologies). Following first strand cDNA synthesis, second strand cDNA was synthesised using 10 units of Escherichia coli polymerase I, 10 units of E. coli DNA ligase and 2 units of RNase H in a reaction containing 25 mM Tris-HCl (pH 7.5), 100 mM KCl, 5 mM MgCl2, 10mM (NH4)SO4, 0.15 mM b-NAD+ and 10 mM dNTPs. The second strand synthesis reaction proceeded at 16 °C for 2 h before 10 units of T4 DNA polymerase was added and the reaction allowed to proceeded for a further 5 minutes. The reaction was terminated by adding 0.5 M EDTA. Double stranded cDNA products were purified using the GeneChip® Sample Cleanup Module (Affymetrix). The synthesised cDNAs were in-vitro transcribed by T7 RNA polymerase (Enzo BioArray High Yield RNA Transcript Labeling Kit, Enzo Life Sciences Inc., Farmingdale, NY, USA) using biotinylated nucleotides to generate biotinylated complementary RNAs (cRNAs). The cRNAs were purified using the GeneChip® Sample Cleanup Module (Affymetrix). The cRNAs were then randomly fragmented at 94 °C for 35 minutes in a buffer containing 40 mM Tris-acetate (pH 8.1), 100 mM potassium acetate, and 30 mM magnesium acetate to generate molecules of approximately 35 to 200 bp. Affymetrix *A. thaliana* ATH1-121501 GeneChip® arrays were hybridised with 15 μg of fragmented labelled cRNA for 16 h at 45 °C as described in the Affymetrix Technical Analysis Manual. GeneChip® microarrays were stained with Streptavidin-Phycoerythrin solution and scanned with an Affymetrix G2500A GeneArray scanner.

Microarray Analysis Suite (MAS Version 5.0; Affymetrix) was used to generate .cel files for each of the biological RNA replicates by scanning and computing summary intensities for each probe without the use of probe mask files. These .cel files were loaded into GeneSpring (Silicon

Genetics, CA, USA) analysis software package using the Robust Multichip Average (RMA) pre-normalisation algorithm (Irizarry et al., 2003). During .cel file loading and pre-normalisation, .cel files were interpreted using either, (1) the *A. thaliana* .cdf file (i.e. with no probe selection used), or (2) using .cdf files generated from the DNA .cel file with DNA hybridisation intensity thresholds from 0 to 1000. Following RMA pre-normalisation and masking of individual probes, further normalisations were applied to the probe set raw signal value in GeneSpring. For each replicate array, each probe set signal value from treated (P-starved) samples was standardised to the probe set signal value of its corresponding control sample. Genes with differential hybridisation intensities between P-starved and control samples were identified using a one-sample, two-tailed Student's t-test, to test whether the mean hybridisation intensity for the gene was significantly different from 1.0 in each data set. Genes with P-values less than 0.05 were considered to be differentially regulated under P starvation.